1П

УДК 681.32

І.О. Погонець
Карпатський державний центр інформаційних засобів і технологій НАН України,
м. Івано-Франківськ, Україна
pogonets@rambler.ru

Теорія ентропії джерел інформації та її застосування в задачах штучного інтелекту

У статті подано теоретичні основи цифрової обробки даних на базі інформаційних мір ентропії для розв'язання задач штучного інтелекту. Проаналізовано існуючі міри ентропії та проведено аналіз статистичних, кореляційних і спектральних характеристик джерел інформації.

Вступ

Розвиток глобальних інформаційних систем в сучасному суспільстві обумовлений швидкими темпами зростання потужних комп'ютерних засобів та телекомунікаційних систем, обумовлює відповідний розвиток засобів та задач штучного інтелекту, які охоплюють проблеми розпізнавання образів, удосконалення систем прийняття рішень, розробку теорії та технології побудови бази знань та ін.

В даній концепції розвитку теорії та методології штучного інтелекту важливу роль відіграє теорія інформації, а особливо її новітня галузь – теорія джерел інформації (ДІ). При цьому фундаментальними основами теорії джерел інформації є теорія ймовірностей, теорія випадкових процесів, теорія кореляційного та спектрального аналізу і особливо теоретичні основи оцінки ентропії.

Важливою характеристикою ентропійного підходу до дослідження джерел інформації, як показано в [1], є найбільш інтегрована оцінка характеристик джерел інформації. Така властивість найбільш інтегрального представлення джерел інформації потребує глибокого дослідження та порівняльного аналізу існуючих мір ентропії, а також диференціації можливостей їхнього застосування при розв'язанні різних задач штучного інтелекту.

1. Аналіз статистичних, кореляційних та спектральних характеристик джерел інформації

Джерело інформації адекватно може бути описано кореляційними, статистичними та спектральними характеристиками [2].

Таблиця 1

Кореляційні моделі		
1	знакова	$B_{xx}(j)=\frac{1}{n}\sum_{i=1}^{n} sign\,\overset{o}{x}_i \cdot sign\,\overset{o}{x}_{i+j}$ $sign\,\overset{o}{x}_i=\begin{cases}+1, \overset{o}{x}_i \ge 0\\ -1, \overset{o}{x}_i < 0\end{cases};$

Продовж. табл. 1

2	релейна	$H_{xx}(j)=\frac{1}{n}\sum_{i=1}^{n}\overset{o}{x}_i\cdot sign\,\overset{o}{x}_{i+j}$
3	коваріаційна	$K_{xx}(j)=\frac{1}{n}\sum_{i=1}^{n}x_i\cdot x_{i+j}$
4	кореляційна	$R_{xx}(j)=\frac{1}{n}\sum_{i=1}^{n}\overset{o}{x}_i\cdot\overset{o}{x}_{i+j}$
5	нормована кореляційна	$\rho_{xx}(j)=\frac{R_{xx}(j)}{D_{xx}}$
6	структурна	$C_{xx}(j)=\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-x_{i+j})^2$
7	модульна	$G_{xx}(j)=\frac{1}{n}\cdot\sum_{i=1}^{n}\lvert x_i-x_{i+j}\rvert$
8	нормована модульна	$g_{xx}(j)=\frac{C_{xx}(j)}{M_x}-M_x$
9	еквівалентна	$F_{xx}(j)=\frac{1}{n}\cdot\sum_{i=1}^{n}\overset{\vee}{Z}_{ij}$; $\overset{\vee}{Z}_{ij}=\begin{cases}x_i, & x_i<x_{i+j}\\ x_j, & x_i\geq x_{i+j}\end{cases}$.
Статистичні моделі		
1	вибіркове математичне сподівання	$M_x=\frac{1}{n}\sum_{i=1}^{n}x_i$
2	ковзне математичне сподівання	$M_j=\frac{1}{m}\sum_{i=1+j}^{m+j}X_{i+j}$, $j=0,1,2,\ldots$ де $j=0,1,2,\ldots$ – дискретний зсув;
3	вагове математичне сподівання	$M_v=\sum_{i=1+j}^{m+j}V_{i-j}\cdot X_{i+j}$, де V_i – вагова функція;
4	дисперсія	$D_x=\frac{1}{n}\sum_{i=1}^{n}(X_i-M_x)^2$
5	середньоквадратичне відхилення	$\sigma_x=\sqrt{D_x}$

2. Теоретичні основи оцінки ентропії

Залежно від динаміки зміни інформаційних станів ДІ класифікують на три типи: 1 – стаціонарні; 2 – квазістаціонарні; 3 – нестаціонарні.

Реальні ДІ найчастіше належать до класу квазістаціонарних джерел, ентропійні моделі яких досліджувалися при вирішенні завдань адаптації за станами, прогнозу випадкових процесів і ефективного кодування. Стаціонарність динамічних і кореляційних показників, стабільність швидкості створення повідомлень при виконанні об'єктами різних технологічних і функціональних операцій визначає тимчасові параметри квазістаціонарних ДІ. Для таких джерел характерним є стрибкоподібний перехід з одного інформаційного стану в інший.

В якості практичної міри ентропії дискретного джерела інформації Р. Хартлі запропонував функцію логарифма числа можливих станів ДІ [3]

$$H = \log \cdot S^n = n \log S,$$

де H – кількість інформації; S – число незалежних рівноймовірних станів ДІ; n – число вибірок.

Інформаційна міра Хартлі не враховує нерівноймовірності розподілу різних S_j-станів. Тому міра Хартлі є верхньою оцінкою ентропії джерела.

При кодуванні безперервних ДІ із заданою точністю квантування за рівнем E і кроком дискретизації за часом А.Н. Колмогоровим запропонована міра інформації – епсилон – ентропія, яка визначається числом елементів E-мережі під час переходу ДІ в різні стани [4]

$$H_E(F) \le \frac{T}{\Delta t} + \log \frac{C}{E},$$

де $\Delta t = \frac{E}{L}$; C – діапазон квантування; T – інтервал часу спостереження ДІ.

Визначивши число функцій, яке може бути отримане у F-просторі за час T у вигляді $\varphi(t) = 2^{\frac{T}{\Delta t}}$, отримаємо

$$H_E \le \log_2 (\frac{C}{E} \cdot 2^{\frac{T}{\Delta t}}).$$

У окремому випадку, коли $\frac{C}{E} = 2^m$ і $\frac{T}{\Delta t} = 2^n$

$$H_E = \log_2 (2^m \cdot 2^n) = m + n,$$

тобто отримуємо ентропію двійково-кодованого ДІ.

Оцінки ентропії ДІ у вигляді міри Хартлі і ентропії Колмогорова вирішувані в цілих числах в тому випадку, якщо діапазон квантування станів ДІ вибирається кратним цілому степеню числа два. У інших випадках, коли $S \ne 2^K$ (K=1,2,..), необхідно користуватися оцінкою

$$H_n = n \cdot \hat{E}[\log_2 S] = n]S[,$$

де E – символ цілочисельної функції з округленням до більшого; $]\cdot[$ – ознака операції округлення до більшого цілого.

Наведені оцінки ентропії ДІ засновані на умові, що кожний S_j стан джерела кодується розрядним двійковим кодом однакової довжини. Розглянуті оцінки ентропії відповідають ДІ з рівноймовірними станами і, як правило, є максимальними.

Для ДІ з нерівноімовірними станами К. Шенноном введена міра ентропії [5]

$$H = -K\sum_{i=1}^{n} P_i \log P_i ,$$

де K – позитивна постійна, яка враховує основу логарифма; P_i – вірогідність S_i-го стану дискретного ДІ.

Б. Олівер відзначає, що чим більше кореляція між наступними один за одним символами або відліками повідомлень, що генеруються ДІ, тим більше нерівномірні розподіли і це веде до зменшення ентропії [6].

Д. Міддлтон також досліджував дискретні ДІ, що формують послідовності символів довільної тривалості, розподілених в певному порядку в часі. Для реалізації дискретної випадкової послідовності $X = \{x_i\}$ кожен з символів x_i може приймати одне із l_i різних значень $(1 \le l_i \le L, \quad i = 1,\ldots,n)$, отриманий вираз для апріорної невизначеності щодо послідовності $(x_1,\ldots,x_i,\ldots,x_n)$

$$H(X) = -\sum_{l_1=1}\cdots\sum_{l_n=1} P(X)\log P(X),$$

де підсумовування проводиться за всіма можливими значеннями кожного з символів x_i.

Для даного джерела визначено вираз середньої умовної ентропії

$$H\left(\frac{X}{Y}\right) = -\sum_{l_1=1}^{L}\cdots\sum_{l_1=1}^{L}\ \sum_{m_1=1}^{M}\cdots\sum_{m_m=1}^{M} P(x_1,\ldots,x_n;\quad y_1,\ldots,y_m) \cdot \log P(x_1,\ldots,x_n;\quad y_1,\ldots,y_m),$$

де x_i, y_i – статистично залежні стани ДІ. З останнього виразу виходить, що для розрахунку ентропії таких ДІ потрібне знання сумісної щільності імовірності різного порядку.

На практиці ДІ не є так статистично складними, щоб їх описувати багатовимірними розподілами. Зокрема для повного опису стаціонарних ДІ достатньо знання двовимірних розподілів і відповідних статистичних середніх. Природно, що ентропія і швидкість створення повідомлень такими джерелами, за рахунок кореляційних зв'язків між різними послідовностями символів, виявляються меншими в порівнянні з оцінками інформаційної міри Хартлі та Шеннона.

У роботі В. Таллера підкреслюються переваги, які можна отримати шляхом кореляційного аналізу і усунення внутрішньої кореляції повідомлень, що формуються джерелом. Показано, що якщо в деякий момент часу ДІ, що має S станів, переходить лише в S_j можливих станів, то аналіз істинності інформаційного змісту його повідомлень приводить до меншого об'єму інформації в порівнянні з функцією найбільшої інформації

$$H \le k2BT\left(1 + \frac{S}{N}\right),$$

де $H = kn \cdot \log Save$, Save – середнє значення станів ДІ; BT – інформаційна база формованих повідомлень; N – значення шуму.

При цьому встановлення вимоги найбільшої інформації засноване на твердженні, що неможливе аналітичне продовження функції інформації з точністю більшою чим $\frac{1}{S}$ впродовж інтервалу між відліками.

Таким чином, в даній роботі В. Таллера практично було поставлено завдання аналізу квазістаціонарних ДІ, а величина $\frac{1}{S}$ отримує смисл інтервалу кореляції між відліками, відповідними станам ДІ.

Численними дослідженнями реальних об'єктів управління і ДІ показано, що параметри технологічних об'єктів на локальних відрізках часу досить точно описуються моделлю двовимірного гауссівського розподілу ймовірностей.

Тому для оцінки сумісної диференціальної ентропії таких джерел скористаємося виразом [5]

$$h(X(t), X(t+\tau)) = \log_2\left(2\pi e \delta_x^2 \sqrt{1-\rho_{xx}^2}\right),$$

який для дискретних ДІ має вигляд

$$H(x_i, x_j) = \log_2 2\pi e + \log_2 \delta_x^2 \sqrt{1-\rho_{xx}^2}\ ,$$

де перший елемент представляє константу інформаційної міри, пов'язану з основою логарифма, яка може не враховуватися в обчисленнях, а другий елемент представляє дисперсію процесу і взаємну ентропію нерівноімовірних корельованих станів ДІ.

На інтервалі кореляції $j = \tau$, який визначається з умови $\rho_{xx}(j) \le 0{,}1$, оцінку ентропії таких ДІ можна визначити за формулою

$$I_{ш} = H(x_i\ , x_j) = \frac{1}{\tau}\sum_{j=0}^{\tau} \log_2 \sqrt{Dx^2 - R_{xx}^2(j)}\ .$$

Розрахунок ентропії ДІ на основі нормованої автокореляційної функції $\rho_{xx}(j)$ є незручним в обчислювальному плані у зв'язку з необхідністю центрування послідовності відліків x_i.

Простіше обчислюється модульна функція автокореляції $G_{xx}(j)$, яка має аналітичний зв'язок з функцією $\rho_{xx}(j)$ у вигляді

$$G_{xx}(j) = 2\sqrt{\frac{\left(\delta_x^2 - R_{xx}(j)\right)}{\pi}},$$

де

$$R_{xx}(j) = \delta_x^2\, \rho_{xx}(j),$$

звідки

$$\rho_{xx}(j) = \frac{\pi}{4} q_{xx}^2(j) - 1,$$

причому

$$q_{xx}(j) = \frac{G_{xx}(j)}{\delta_x}$$

нормована модульна функція автокореляції.

Після нескладних перетворень оцінку ентропії отримаємо у вигляді

$$I_{uu} = \frac{1}{\tau}\sum_{j=0}^{\tau}\log\frac{\pi e\sqrt{\pi}}{2} \quad G_{xx}(j)\sqrt{8-\pi q_{xx}^{2}(j)}\ ,$$

яка аналогічним чином приводиться до трьох елементів

$$I_{uu} = \log_2\frac{\pi e\sqrt{\pi}}{2} + \frac{1}{\tau}\sum_{j=0}^{\tau}\log_2 G_{xx}(j) + \frac{1}{2\tau}\sum_{j=0}^{\tau}\log_2\left[8-\pi q_{xx}^{2}(j)\right],$$

де другий елемент представляє число – випадкову складову еквівалентну епсилон ентропії, а третій – взаємну ентропію корельованих відліків.

Серед розглянутих інформаційних мір ентропії найбільш повно опирається на статистичні характеристики міра ентропії на базі кореляції.

Література

1. Погонець І.О., Николайчук Я.М. Методи визначення ентропії джерел інформації // Вісник Хмельницького національного університету. – Хмельницький: ХНУ. – 2007. – Т. 1(90), № 2. – С. 93-99.
2. Пітух І. Кореляційні та ентропійні моделі об’єктів управління розподілених комп’ютерних мереж. – Івано-Франківськ: Наукові вісті, Інститут менеджменту та економіки «Галицька академія». – 2006. – № 2(10).
3. Хартли Р.Л. Передача информации // Теория информации и ее приложения. – М., 1959. – 350 с.
4. Колмогоров А.Н. Теория передачи информации: Сес. АН СССР по науч. пробл. автоматизации пр-ва. Пленар. заседания. – М.: Изд-во АН СССР, 1957. – 15с.
5. Шеннон К.Э. Работы по теории информации и кибернетика. – М.: Изд-во иностр. лит., 1963. – 829 с.
6. Оливер Б. Эффективное кодирование // Теория информации и ее приложения. – М., 1959. – С. 1-15.

И.О. Погонець

Теория энтропии источников информации и ее приложение в задачах искусственного интеллекта

В статье даны теоретические основы цифровой обработки данных на базе информационных мер энтропии для решения задач искусственного интеллекта. Проанализированы существующие меры энтропии и проведен анализ статистических, корреляционных и спектральных характеристик источников информации.

Стаття надійшла до редакції 10.07.2008.